Emergent Mental Lexicon Functions in ChatGPT

Christopher Kello (ckello@ucmerced.edu) and Polyphony Bruna (pbruna@ucmerced.edu)
Cognitive and Information Sciences, University of California, Merced
5200 North Lake Rd., Merced, CA 95343 USA

Abstract

Traditional theories of the human mental lexicon posit dedicated mechanisms of processing that develop as sustained functions of brain and mind. Large Language Models (LLMs) provide a new approach in which lexical functions emerge from the learning and processing of sequences in contexts. We prompted lexical functions in ChatGPT and compared numeric responses with averaged human data for a sample of 390 words for a range of lexical variables, some derived from corpus analyses and some from Likert ratings. ChatGPT responses were moderately to highly correlated with mean values, more so for GPT-4 versus GPT-3.5, and responses were sensitive to context and human inter-rater reliability. We argue that responses were not recalled from memorized training data but were instead *soft-assembled* from more general-purpose representations. Emergent functions in LLMs offer a new approach to modeling language and cognitive processes.

Keywords: large language models; emergent functions; mental lexicon; soft-assembly

LLMs have developed a surprising and impressive ability to perform a wide range of prompted tasks on which they were not directly trained (Wei, Tay, et al., 2022). Training is primarily based on learning sequences through prediction in the context of surrounding tokens (e.g. words), and training data consists of many long sequences of text and symbols culled from various corpora. After training, a prompt can be written to begin a context, and then an LLM like that used in ChatGPT can continue and expand the context by autoregressively predicting and producing subsequent tokens until an end of response is predicted. As a result, LLMs can generalize from their training data to produce responses to prompts that answer questions, perform tasks, solve problems, and generate plausible text and/or numbers given the context (OpenAI, 2023). Continuations can be sometimes bizarre, offensive, biased, untruthful, or otherwise unhelpful, so LLM responses are also fine-tuned to be more helpful and appropriate, often through reinforcement learning based on targeted sets of prompts and responses (Ouyang et al., 2022).

LLM functions assembled from prompts are often referred to as *emergent* because they are untrained and become available as training corpora and model parameters become larger (Bommasani et al., 2021). Fine-tuning plays a role in shaping LLM responses, especially those targeted by human feedback, but this feedback is limited relative to the remarkable breadth and scope of emergent functionality.

Investigations are underway to understand how LLM functions emerge from transformer networks. One technique is to prompt them with examples of the desired function so the LLM may generalize by similarity and analogy (Wei, Wang, et al., 2022). Such *few shot* learning protocols can be effective, but functions can also be assembled with no examples, so-called *zero shot* learning or reasoning (Kojima et al., 2022). Few and zero shot learning suggests that transformer networks learn general-purpose representations that serve as pre-trained "building blocks" ready for on-the-fly assembly directed by prompting (Jiang & Bansal, 2021).

Emergent function stands in contrast with the idea that prompts serve as cues for locating and recalling functions in training data memorized in ChatGPT's weight matrices (Reynolds & McDonell, 2021). Indeed, LLMs can memorize sequences from their training data (Carlini et al., 2022), and the function of memorization may be as emergent as any other LLM function (Biderman et al., 2023). We aim to test emergence versus memorization as a window into how LLMs may serve as models of human language and cognition.

The Mental Lexicon

We investigate the emergence versus memorization hypotheses by prompting ChatGPT with tests of the human *mental lexicon*, which refers to a person's knowledge of, and facility with, the words of their language, including spelling, sound, meaning, and grammar (Aitchison, 2012). The mental lexicon has long been a testbed for theoretical frameworks of cognition and information processing, most notably in debates between rule-based versus connectionist processing (Coltheart et al., 1993; Smolensky, 1988).

LLMs are founded on connectionist processing, but they offer a different theory of the human mental lexicon, and of human cognition more generally, compared with prior connectionist or rule-based theories. For instance, the Dual-Route Cascade model (Coltheart et al., 2001) of word and nonword reading posits several rule-based mechanisms for mapping letters to sounds, accessing the stored meanings and sounds of words, and producing sequences of phonemes. By contrast, the distributed connectionist (Plaut et al., 1996; Seidenberg & McClelland, 1987) posits learned mappings between the spellings, sounds, and meanings of words.

The theories espouse different principles of word and nonword reading, but they make the same implicit assumption that the posited mechanisms and mappings are *dedicated* functions of *sustained* structures in the brains and minds of readers. Model outputs can be shaped by task demands and context if enabled by the mechanisms or mappings, but the model architectures cannot reorganize themselves for different purposes under different conditions. Assumptions of dedicated functions and sustained structures are held by most models of language of cognitive functions.

LLMs offer a different approach to theorizing about the mental lexicon, and language and cognition in general.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Rather than dedicated functions or sustained structures, lexical processing emerges as a *potential* in learning a generative model of language. The potential is realized in prompted contexts and task conditions. Simple Recurrent Networks provided an early glimpse into this emergent approach in that they showed how lexical processes could emerge from learning sequences of linguistic units (Elman, 2005; Sibley et al., 2008). LLMs greatly expand on the emergent approach by enabling the assembly of potentially any function of the mental lexicon, as just one of many domains of function assembly enabled by sequence learning in transformer networks.

We used zero shot prompts in ChatGPT to elicit several lexical functions whose outputs are numeric and can be compared with corpus statistics and human word ratings. As LLM responses more closely match the data, functions come closer to constituting a model of the human mental lexicon created by sequence learning in LLMs. Mental lexicon models are often tested against measures of online lexical processing such as response times or event-related potentials (ERPs), but LLM processing is generally not theorized to simulate mental or neural processing dynamics in humans. Therefore, only lexical statistics and ratings are tested herein.

Experimenting with Lexical Functions in LLMs

We conducted experiments with ChatGPT designed to test emergence versus memorization as the basis of its ability to perform lexical functions. Words and associated data came from the South Carolina (SCOPE) Psycholinguistic metabase (Gao et al., 2023). We chose SCOPE *corpus variables* that are known to correlate with measures of human lexical processing such as response times and error rates. This choice allowed us to prompt ChatGPT for corpus values as a proxy for online measures of human lexical processing. We also chose *rating variables*, whose values could be prompted from ChatGPT as they were prompted from participants. We focused on the Glasgow norms (Scott et al., 2019) because they tap into a range of lexical variables, from a word's age of acquisition and familiarity to its arousal and dominance. We added two additional rating variables from other studies to further expand the range and test outside the Glasgow study for comparison.

We tested and compared GPT-3.5 responses versus GPT-4 responses in chat mode, and we expected GPT-4 responses to more closely match corpus statistics and human ratings compared with GPT-3.5. GPT-4 is reported to be much larger in terms of model parameters and training data, with more fine-tuning of its responses. A given lexical function may be effectively absent in GPT-3.5 and present in GPT-4, or present in both models but more human-like in GPT-4.

Both emergence and memorization lead us to expect more human-like responses from GPT-4, but for different reasons. On emergence, lexical functions are assembled from general-purpose representations (Bender et al., 2021), and these learned representations would better reflect patterns of human behavior that are better captured by larger models trained on larger corpora. On memorization, larger training corpora would be more likely to explicitly include the values of lexical variables, and larger models would be better able to recall those values from its weight matrices (Bender et al., 2021).

The emergent and memorization hypotheses diverge, however, when one considers correlations with specific lexicon variables. Memorization predicts a comparable degree of correlation for different variables from the same study—if the LLM can recall ratings for one variable from the Glasgow norms, it should similarly be able to recall ratings for other Glasgow variables. It is possible that some variables may serve as better cues to recall than others, but a process akin to random access memory would either recall values as memorized, or not. Approximated values would necessarily come from an assembled process that involves more general knowledge about the meanings of variables, usages of words, and how to map variables and words onto values. Therefore, we argue that different degrees of correlation for different variables from the same study would stand as evidence against the memorization hypothesis.

In contrast, the emergent hypothesis allows for variability in how each lexicon function is assembled depending on the lexical variable and words being probed. This variability is context-dependent (Kello et al., 2007), which means that more ambiguous lexicon variables and words may lead to more variability across contexts and hence more varied response values. Human ratings should also be more varied if the human mental lexicon is similarly affected by ambiguity. Therefore, the emergent hypothesis predicts that the strength of LLM correlations will correspond with the reliability of human responses. We can test this prediction because the Glasgow study includes not only mean word ratings for each lexicon variable, but also the standard deviation (SD) across individual raters for each word and variable, i.e. measures of inter-rater reliability. The emergent hypothesis predicts lower LLM correlations for words and lexicon variables with lower human inter-rater reliability.

Experiment 1

The SCOPE corpus currently contains more than 250 variables and over 100,000 words and 81,000 nonwords gathered from dozens of psycholinguistic studies. Any given word may have data for only a subset of the variables, so we selected all and only the words that had data for all variables chosen for inclusion in this study. We chose 4 variables derived from corpus statistics and 12 variables of mean human word ratings for a total of 16 variables and 390 words (the selected words were also required to include data on 5 additional variables chosen for future study). The variables are listed in Table 1 and the words can be found at https://github.com/pjbruna/chatgpt-soft-assembly along with prompts and data collection and analysis scripts.

Each prompt was designed to elicit estimated values for just one of the 16 SCOPE variables for a given set of words, and ChatGPT was prompted separately for each variable. The GPT models used through the OpenAI API were `gpt-4-0613` and `gpt-3.5-turbo-16k-0613`. Preliminary

testing showed that GPT-3.5 was not always able to provide valid responses for large sets of words, so for both models, we randomly divided the selected sample of 390 words into 5 batches of 78 words each, with words ordered randomly.

Table 1: SCOPE Sources, Variables, and Value Types

Source	Variable	Type
Brysbaert & New (2009), corpus	Word Frequency	Int [1, 1M]
	Contextual Diversity	Int [1, 1000]
Hoffman et al. (2013), corpus	Semantic Diversity	Real [0, 2.5]
Webster's Dict.	Num. Meanings	Int [1, *N*]
Kuperman et al. (2013)	Age of Acquisition	Int [1, *N*], ratings
Scott et al. (2019) Glasgow norms	Age of Acquisition	Real [1, 7], ratings
	Concreteness	
	Familiarity	
	Gender Association	
	Imageability	
	Semantic Size	
	Arousal	Real [1, 9], ratings
	Dominance	
	Valence	
Diveica et al. (2022)	Socialness	Real [1, 7], ratings
Engelthaler & Hills (2018)	Humor	Real [1, 5], ratings

Prompts for word ratings closely followed the prompts given to human raters. Small edits were made to adjust for responding with a set of ratings at a time, rather than individual words and ratings. Prompts for the corpus variables directed the LLM to estimate the statistic based on the method by which it was computed. For instance, word frequency and contextual diversity were based on a corpus of film subtitles, so the prompts asked for estimates based on an "imagined" corpus of film subtitles. The imagined size was changed to indicate a standardized 1,000 films with 1M words, and responses were scaled and logged to match the size of the corpus.

Some of the rating prompts for humans made references to the rater in a personal tone of voice. For ChatGPT, the tone of the prompt tends to evoke a similar tone of response, and preliminary tests indicated that such a tone may increase the chance of some embellished narrative being included with responses for which only words and ratings are requested. To set a formal, task-oriented tone, ChatGPT prompts for both human ratings and corpus statistics were written in a directive register, mostly using the imperative tense.

A separate ChatGPT user prompt was created for each of the 16 SCOPE variables, followed by each of the 5 batches of words. In addition to the user prompt, ChatGPT includes a system prompt that sets a prior, overarching context for a given chat interaction. We wrote a single system prompt as context for all user prompts that directed the LLM to encode the upcoming task and respond with words and values in a specific format. The system prompt and example user prompts (excluding words) are shown in Table 2.

Table 2: System Prompt and 3 Example User Prompts

System Prompt: "*Encode the task and then for each and every word listed after the task instructions, respond with its corresponding value, one 'word,value' pair per line.*"
Contextual Diversity: "*The contextual diversity of a word is based on the number of different contexts that the word appears in. Contextual diversity is measured from the subtitles of films, where each film is a different context, and a word may or may not appear in the subtitles of a given film. Imagine a sample of one thousand films with American English subtitles, and a given word may appear in some number of different films based on their subtitles. Estimate the number of films, from 1 to 1000, that each of the following words appears in:*"
Number of Meanings: "*Dictionaries list all the known meanings of words in a language. For a typical English language dictionary, provide a whole number estimate of the number of meanings for each of the following words:*"
Dominance: "*Dominance is a measure of the degree of control felt by a person. A word can make a person feel DOMINANT, influential, in control, important, or autonomous. In contrast, a word can make a person feel CONTROLLED, influenced, cared-for, submissive, or guided. Indicate how each word makes a person feel on a continuous scale of 1.0 (VERY CONTROLLED) to 9.0 (VERY DOMINANT), with the midpoint being neither controlled nor dominant:*"

ChatGPT enables users to also set the size of the context window, in number of tokens, used by the LLM transformer network and the temperature parameter on the autoregressive next-token prediction output layer. We set the window size to be large enough to encompass each system and user prompt and its ChatGPT response (5,000 tokens), and we set the temperature to zero, which means that the maximally activated token is always chosen on each step of processing. ChatGPT has one or more other sources of randomness not visible or accessible to the user, so ChatGPT responses are not fully deterministic even with temperature set to zero. We ran each prompt twice to measure the average deviation in response values due to stochasticity. GPT-4.0 always responded with numeric values for all words, GPT-3.5 gave non-numeric (e.g. "?") or anomalous (values well outside the possible range) on a few rare occasions. Another prompting was sufficient in these cases to yield valid responses.

Results

Analyses were performed in the R environment (v. 4.2.2), using the `tidyverse` (Wickham et al., 2019), `ggplot2` (Villanueva & Chen, 2019), `lmerTest` (Kuznetsova et al., 2017), and `emmeans` (Lenth, 2023) packages. Table 3 presents correlations between SCOPE and GPT values for

each variable and corresponding GPT prompt, for each of two model runs. Each correlation is based on the same 390 words divided into the same 5 randomized batches. The average increase in correlation from GPT-3.5 to GPT-4 is in the right-most column, and rows are sorted by average increase.

Table 3: Correlations between SCOPE and GPT values

	GPT-3.5		GPT-4		
	Run 1	*Run 2*	*Run 1*	*Run 2*	*Avg Inc.*
Imageability	0.03	0.05	0.82	0.79	0.77
Gender	0.28	0.32	0.84	0.86	0.55
Semantic Div.	-0.01	-0.04	0.42	0.51	0.49
AoA Glasgow	0.46	0.52	0.88	0.88	0.39
Frequency	0.22	0.23	0.64	0.58	0.39
AoA Kuper.	0.56	0.55	0.87	0.86	0.31
Semantic Size	0.51	0.47	0.79	0.78	0.30
Socialness	0.59	0.62	0.87	0.87	0.27
Context Div.	0.51	0.42	0.69	0.78	0.27
Concreteness	0.66	0.68	0.92	0.92	0.25
Familiarity	0.50	0.47	0.72	0.72	0.24
Humor	0.36	0.37	0.57	0.56	0.20
Meanings	0.55	0.61	0.66	0.66	0.08
Dominance	0.55	0.55	0.62	0.63	0.08
Valence	0.87	0.88	0.94	0.94	0.07
Arousal	0.48	0.48	0.52	0.50	0.03
MEAN	**0.45**	**0.45**	**0.74**	**0.74**	**0.29**

Table 3 shows little difference between runs, indicating very minor effects of stochasticity at zero temperature. In general, correlations indicate that ChatGPT responses were consistent with corpus values and human responses from SCOPE. That said, there was a wide dispersion in effect sizes across variables and GPT models, from uncorrelated to nearly perfectly correlated. GPT-4 responses were substantially more correlated with SCOPE data than GPT-3.5 responses for most variables, and there was also wide dispersion in the jumps in correlation, from no appreciable increase to a full jump from uncorrelated to highly correlated.

The wide dispersion in correlations across variables in the SCOPE corpus is evidence that lexical functions were assembled rather than recalled from memory. Wide dispersion in the size of correlation jumps is evidence that the emergence of lexical functions in GPT models is a heterogeneous process—jumps in correlation for some functions were discontinuous like a phase transition, jumps for others were more gradual, and still others showed no detectable change in correlation.

To test whether heterogeneity in emergence may be partly explained by variation in the ambiguity of lexical variables, we averaged the human item-level standard deviations (SDs) for each Glasgow variable, and we plotted their relationship with correlations between corresponding mean ratings and ChatGPT responses (Figure 1). Scatter plots show a strong correspondence between SDs and ChatGPT correlations for GPT-4, but not GPT-3.5 (r = -0.81 and -0.03, respectively, using the Fisher Z transform on ChatGPT correlations). To test whether this effect might instead come from variation in Glasgow means (i.e. an effect of restricted range), we removed variability in item SDs that was accounted for by variability in mean SDs. The result held up, albeit slightly weakened (r = -0.63 and 0.10 for GPT-4 and GPT-3.5).

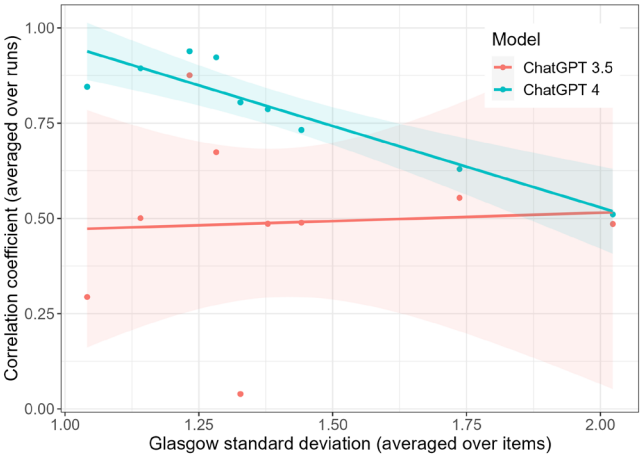

Figure 1: Relationship between ChatGPT correlations and inter-rater reliability for the Glasgow norms

The above results show that assembly of lexical functions in GPT-4 was better for variables with greater human inter-rater reliability, as measured by mean item SDs. We tested the same effect at the individual word level by pairing words that had equivalent mean ratings for a given Glasgow variable but differed by at least 0.1 in their item SDs, thereby creating high versus low SD word groups for each variable. Correlations with mean Glasgow ratings for GPT-3.5 and GPT-4 are shown in Table 4, for high and low SD groups. Differences in correlations between high and low SD groups are also shown, where positive numbers are in the predicted direction of stronger correlations for words with greater inter-rater reliability.

Table 4: Correlations between Glasgow and GPT for Words with High vs Low Item SDs

	GPT-3.5			GPT-4		
	High SD	*Low SD*	*Diff*	*High SD*	*Low SD*	*Diff*
AoA	0.49	0.47	−.02	0.88	0.90	+.02
Arousal	0.49	0.46	−.03	0.47	0.55	+.08
Concrete	0.66	0.64	−.02	0.93	0.90	−.03
Dominant	0.53	0.60	+.07	0.59	0.69	+.10
Familiar	0.51	0.50	−.01	0.70	0.74	+.05
Gender	0.30	0.34	+.04	0.81	0.87	+.07
Imageable	.007	0.05	+.04	0.79	0.80	.004
Size	0.48	0.53	+.05	0.77	0.81	+.04
Valence	0.86	0.88	+.02	0.92	0.95	+.02
MEAN	**0.48**	**0.50**	**+.01**	**0.76**	**0.80**	**+.04**

Results at the word level of analysis followed the same pattern as results at the variable level: The assembly of lexical functions in GPT-4 was better for words with greater human inter-rater reliability, as measured by mean item SDs, whereas this effect was less clear for GPT-3.5. We tested the effect of reliability (high versus low item SD) using a one-sample paired t-test on the differences for each GPT model, with the Fisher *Z* transform of correlations as the dependent measure. We found that correlations for the low SD words were significantly higher than those for high SD words for GPT-4 ($t(8) = 2.44$, $p = 0.02$) but only marginally for GPT-3.5 ($t(8) = 1.29$, $p = 0.12$). In conclusion, function assembly in GPT-4 is sensitive to inter-rater reliability, which we interpret as further evidence for the emergence of a model of the human mental lexicon in GPT-4.

Experiment 2

Experiment 1 showed that ChatGPT can be prompted to assemble lexical functions that closely follow corpus and human response data, more so in GPT-4 compared with GPT-3.5. The ability of ChatGPT to assemble myriad other functions, lexicon and otherwise, suggests that assembled functions will tend to be highly context sensitive. The reason is that even small changes in prompts must have the possibility of causing substantial modifications in the assembled function. For example, how words are judged in terms of arousal could vary based on subtle cues in the prompt, including the context of other words being judged. Such word *list composition* effects are well-documented in human responses (e.g. Dorfman & Glanzer, 1988), and in ChatGPT, they would provide evidence against the existence of a pre-determined function for each lexical variable.

In Experiment 2, we investigate the contextuality of lexical functions by testing for effects of list composition. Words in Experiment 1 were placed in the same 5 batches and listed in the same random order for all prompts. In Experiment 2, we ran the same variable prompts for the same 390 words but in two different arrangements. One was a different randomized order from Experiment 1, resulting in 5 different batches of randomized words. In the other arrangement, words were sorted according to the mean values of each SCOPE variable, and then 5 sorted batches were extracted by taking every fifth word in the sorted lists. Thus, unbeknownst to ChatGPT, each prompt called for responses to be in numeric order, from lowest to highest value.

If assembled lexical functions are sensitive to context, then ChatGPT responses should vary more between different batches and orders compared with two runs of the same batches and orders at zero temperature. As noted in Experiment 1, some variability can be attributed to stochasticity even at zero temperature, so we compared different batches and orders against a baseline of stochastic variability. If the sensitivity to context specifically affects each assembled lexical function, then the sorted condition should vary more from baseline compared with the random condition because words were sorted according to the SCOPE variable for each respective prompt.

Results

To measure the degree of GPT variability, we calculated the absolute difference between ratings for each word, for both Run 1 and 2 in Experiment 1, and the two new runs: the newly randomized (*Rnd*) and sorted (*Srt*) word orders. The absolute difference between the two runs from Experiment 1 was used as a control (*Crl*).

Table 5 presents the mean absolute difference for each SCOPE variable, averaged over the Run 1 and 2 baselines for *Rnd* and *Srt* conditions. As predicted, the table shows greater differences for different batches and word orders (*Rnd* and *Srt*) compared with the baseline of ChatGPT stochastic variability (*Crl*). The differences also appear greater for *Srt* vs. *Rnd*, and greater for GPT-3.5 vs. GPT-4.

Table 5: Average absolute differences in Experiment 2

	GPT-3.5			**GPT-4**		
	Crl	*Rnd*	*Srt*	*Crl*	*Rnd*	*Srt*
Imageable	0.12	0.73	0.70	0.17	0.52	0.77
Gender	0.19	0.67	0.80	0.09	0.45	0.86
Sem. Div	0.12	0.60	0.52	0.28	0.37	0.45
AoA Glas.	0.16	0.66	0.73	0.11	0.62	0.67
Frequency	0.77	1.07	1.02	0.55	0.75	0.40
AoA Kup.	0.30	1.66	2.68	0.23	1.13	1.39
Sem. Size	0.06	0.71	0.92	0.16	0.57	0.77
Social	0.28	0.87	1.19	0.14	0.60	0.63
Context Div	0.32	0.31	0.49	0.15	0.20	0.20
Concrete	0.13	0.59	0.68	0.25	0.45	0.60
Familiar	0.10	0.75	1.03	0.31	0.44	0.65
Humor	0.13	0.38	0.43	0.10	0.44	0.57
Meanings	0.23	1.49	0.99	0.06	0.57	0.57
Dominance	0.09	0.76	0.92	0.26	0.78	0.99
Valence	0.11	0.56	0.77	0.08	0.48	0.91
Arousal	0.06	0.78	1.62	0.30	0.77	0.67
MEAN	**0.20**	**0.79**	**0.97**	**0.20**	**0.57**	**0.69**

We tested the apparent effects using a linear mixed-effects regression model with comparison type (*Crl/Rnd/Srt*) and GPT model (3.5/4) as fixed effects, and SCOPE variable and word as random effects. Here we report results with Run 1 as the baseline (these findings replicated when Run 2 was used instead). We found a main effect of comparison type but not model type. Averaging over GPT models, both *Rnd* and *Srt* differences were greater than the *Crl* stochastic baseline ($b = 0.59$, $p < 2e\text{-}16$ for *Rnd*; $b = 0.77$, $p < 2e\text{-}16$ for *Srt*), and *Srt* differences were greater than *Rnd* differences ($b = 0.15$, $p < 0.0001$).

In addition, the means in Table 5 suggest that context effects may have been stronger for GPT-3.5 vs. GPT-4. In support of this apparent interaction, we ran pairwise comparisons and found that *Rnd* and *Srt* differences were greater for GPT-3.5 compared with GPT-4 ($b = 0.21$, $p < 0.0001$ for *Rnd*; $b = 0.28$, $p < 0.0001$ for *Srt*).

ChatGPT Assembly Consistency. The interaction result suggests that the assembly of lexical functions in GPT-4 is

more robust to context effects compared with GPT-3.5, as if the functions are based on more stable "concepts" underlying the words and lexical variables. We tested this hypothesis by comparing the consistency of GPT responses across the two age of acquisition (AoA) prompts, where one requested age estimates in years, and the other requested ratings on 7-point scale where each number corresponded to an age range. If GPT-4 has a more stable concept of AoA, then its responses should be more self-consistent compared with GPT-3.5. As predicted, Values for the two different AoA response formats were more highly correlated for GPT-4 (r = 0.91) compared with GPT-3.5 (r = 0.55). This result provides evidence for greater consistency across contextual variations in assembled functions, and it calls for further investigation into the notion of lexical and conceptual stability in ChatGPT.

Conclusions and Discussion

Our experiments demonstrate the emergence of a model of the human mental lexicon in ChatGPT, with more human-like lexical functions in GPT-4. Correlations with corpus variables suggest that LLMs can estimate lexical statistics, either directly or indirectly from training corpora, but not by recalling specific values from memory. Correlations with mean human ratings were surprisingly strong for some lexical variables, and for GPT-4, correlations were stronger for Glasgow variables and words with greater human inter-rater reliability. These results further support emergence over memorization and indicate how LLMs can provide a different kind of model of human language and cognition compared with prior connectionist and rule-based models.

First and foremost, an LLM mental lexicon does not exist outside of tasks that trigger the assembly of specific lexical functions in particular contexts (Elman, 2009). The diversity of response correlations across lexical variables and between GPT models indicated that lexical functions were *soft-assembled* (Kello & Van Orden, 2009) from more general-purpose distributed representations learned in the service of next-token prediction. Prompts provided context to guide and constrain the process of soft assembly, which can be highly sensitive to even subtle changes in context, which was evidenced by ChatGPT word list composition effects.

Prior to context, lexical functions only exist as latent potentials, which means that they may manifest in different ways when soft-assembled from prompts that differ even superficially in the requested response format. We found evidence for this sensitivity in the comparison of GPT-3.5 responses for two different response formats for age of acquisition, which showed only a moderate correlation despite using the same word lists. Given that stochasticity inherent to ChatGPT was not sufficient to explain the unaccounted variability, we can conclude it came from differences in the soft-assembled age of acquisition function. GPT-4 showed more consistent function assembly, in that different response formats for age of acquisition were highly correlated with each other.

For mental and neural activity, soft-assembly has been based on a balance of interdependence and independence among the system components from which functions arise (Tognoli & Kelso, 2014). This balance is hypothesized to result from dynamical metastability in neural and behavioral patterns of activity and their associated functions. LLM dynamics appear to be very different from human cognitive dynamics, and LLM models currently are not used to simulate human language learning or the time course of processing. Nevertheless, metastability may be a property of LLMs that arises from learning to predict tokens for an enormous range of trained contexts. In support of this conjecture, Hopfield networks with metastable states were recently incorporated with LLMs (Ramsauer et al., 2020), resulting in flexibility in the degree of assembly versus memorizations.

The emergent nature of LLM functions means that we do not have direct access to an assembled model of the mental lexicon because it is buried in the LLM parameters, along with the activation values triggered by the prompted contexts and responses. Detailed model information is not made available by OpenAI for ChatGPT, but even for open-source LLMs of sufficient size, the weight matrices and unit functions make it daunting to isolate a complex emergent function like the mental lexicon. A promising direction of research on *mechanistic interpretability* seeks to reverse engineer the functionality of LLMs and other large-scale deep learning models (Nanda et al., 2023). Progress on mechanistic interpretability will be critical in using LLMs as models of the human mental lexicon, and language and cognition in general.

In lieu of direct investigations into transformer mechanisms and learned representations, researchers may use experimental methods with LLMs and compare results with human experiments, as we have done herein. Some soft-assembled LLM functions can be compared directly against measures of language and cognition, such as those based on Likert ratings herein. Other functions are abstracted or implemented differently in LLMs and so require more indirect comparisons, such as LLM estimates of corpus statistics as measures of effects on human on-line processing. Foundation models may evolve to provide more direct measures on on-line processing in the future.

Lastly, it is an open question whether LLMs provide models of language and cognition at the level of individuals or populations (Aher et al., 2023; Andreas, 2022). Our prompts elicited ChatGPT responses as if it was an individual, but we compared its responses to sample means and SDs intended to represent a population. Further experiments may inform this issue and others at the intersection of LLMs and human language and cognition.

Acknowledgments

Thank you to the undergraduate research assistants who participated in discussions and presentations of this research, with special thanks to Katelyn Prater and Kanly Thao. Polyphony Bruna was supported by a UC Merced Cognitive and Information Sciences Fellowship.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. International Conference on Machine Learning,

Aitchison, J. (2012). *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons.

Andreas, J. (2022). Language models as agent models. *arXiv preprint arXiv:2212.01681*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency,

Biderman, S., Prashanth, U. S., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., & Raf, E. (2023). Emergent and predictable memorization in large language models. *arXiv preprint arXiv:2304.11158*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., . . . Brunskill, E. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods, 41*(4), 977-990.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 589-608.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204-256.

Diveica, V., Pexman, P. M., & Binney, R. J. (2023). Quantifying social semantics: An inclusive definition of socialness and ratings for 8388 English words. *Behavior Research Methods, 55*(2), 461-473.

Elman, J. L. (2005). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*, 301-306.

Elman, J. L. (2009). On the Meaning of Words and Dinosaur Bones: Lexical Knowledge Without a Lexicon. *Cognitive Science*, *33*(4), 547-582.

Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior Research Methods, 50*, 1116-1124.

Gao, C., Shinkareva, S. V., & Desai, R. H. (2023). Scope: The south carolina psycholinguistic metabase. *Behavior Research Methods*, *55*(6), 2853-2884.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods, 45*, 718-730.

Jiang, Y., & Bansal, M. (2021). Inducing Transformer's Compositional Generalization Ability via Auxiliary Sequence Prediction Tasks. *arXiv preprint arXiv:2109.15256*.

Kello, C. T., Beltz, B. C., Holden, J. G., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of experimental psychology. General*, *136*(4), 551-568.

Kello, C. T., & Van Orden, G. C. (2009). Soft-assembly of sensorimotor function. *Nonlinear Dynamics, Psychology, and Life Sciences*, *13*(1), 57-78.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199-22213.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978-990.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*(13).

Lenth, R. (2023). emmeans: Estimated Marginal Means, aka Least-Squares Means. *R package version 1.8.4-1*.

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

OpenAI. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., . . . Ray, A. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56-115.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., . . . Sandve, G. K. (2020). Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. Extended Abstracts of the *2021 CHI Conference on Human Factors in Computing Systems*.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500

words on nine scales. *Behavior Research Methods*, *51*, 1258-1270.

Seidenberg, M. S., & McClelland, J. L. (1987). A Distributed, Developmental Model of Visual Word Recognition and Naming. *Bulletin of the Psychonomic Society*, *25*(5), 329-329.

Sibley, D. E., Kello, C. T., Plaut, D. C., & Elman, J. L. (2008). Large-Scale Modeling of Wordform Learning and Representation. *Cognitive Science*, *32*(4), 741-754.

Smolensky, P. (1988). On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences*, *11*(1), 1-23.

Tognoli, E., & Kelso, J. A. S. (2014). The Metastable Brain. *Neuron*, *81*(1), 35-48.

Villanueva, R. A. M., & Chen, Z. J. (2019). ggplot2: elegant graphics for data analysis. Taylor & Francis.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., . . . Metzler, D. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., . . . Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., . . . Hester, J. (2019). Welcome to the Tidyverse. *Journal of open source software*, *4*(43), 1686.